知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

我国文献资源保障体系论文主题识别与演化分析

田栩冉[1] 马笑笑[1] 李玉海[1,2]
1. 华中师范大学信息管理学院 武汉 430079
2. 中国图书馆创新发展研究中心 武汉 430079

摘要：[目的/意义]对我国文献资源保障体系论文主题演化的路径进行分析，为重构我国文献资源保障体系提供借鉴。[方法/过程]主要使用 LDA 模型对所搜集到的文献进行聚类，首先以时间段为划分依据，绘制主题词共现网络，探索各主题之间的交互关系，后通过相似度计算判定各主题内部的演化路径并绘制桑基图以可视化形式展现演化结果。[结果/结论]研究发现我国文献资源保障体系的相关主题在 2000 年左右均已基本出现，主题主要包含资源角度和机构角度两大类，且受计算机技术和国家政策影响较大，并针对该两大类主题，给出相应的对策与建议。

关键词：文献资源保障体系 LDA 主题识别 主题演化

分类号：G253

引用格式：田栩冉，马笑笑，李玉海. 我国文献资源保障体系论文主题识别与演化分析 [J/OL]. 知识管理论坛，2021, 6(6): 303-314[引用日期]. http://www.kmf.ac.cn/p/263/.

1 引言

目前全球竞相步入 5G（第五代移动电话行动通信标准，也称第五代移动通信技术）时代，相比之前的 4G 时代，网络数据的传输速度将会更快，5G 技术可以被更快速更高效地运用到多个领域。传统的文献资源与新型的数字文献资源数量不断累积，通过文献信息资源的整体建设，建立起能在一定范围内有效保障社会文献需求的文献信息资源系统——文献资源保障体系[1]。

在这样一个本该互联互通的时代浪潮之下，国外数据库商依仗其丰富的文献资源，坐地起价，企图继续垄断资源，引发了国内图书馆人的不满。欧洲大学协会（European University

基金项目：本文系国家社会科学基金重大课题“新时代我国文献信息资源保障体系重构研究”（项目编号：19ZDA345）研究成果之一。

作者简介：田栩冉，硕士研究生，E-mail：tianxuran@mails.ccnu.edu.cn；马笑笑，硕士研究生；李玉海，华中师范大学信息管理学院院长，教授，博士。

收稿日期：2021-09-01 发表日期：2021-11-01 本文责任编辑：刘远颖

Association，EUA）发布的一份报告显示，学术机构、图书馆与美国化学学会（ACS）、爱思唯尔（Elsevier）、威利（Wiley）、施普林格·自然(Springer Nature)和泰勒弗朗西斯集团(Taylor ＆ Francis）等出版商的交易成本正以每年 3.6% 的速度上涨。文献资源，尤其是科技文献资源，是对科学最新前沿研究结果的展现，如果放弃相关资源的购买，则会丧失国际科研竞争力；如果继续服从霸王条款，依然无法改变被动的局面。故而，在以程焕文先生为代表的《十问数据商 !!!》等一系列诘问之后，重构我国的文献资源保障体系成为当务之急。

目前国外已有一部分高校通过开放获取出版的方式应对数据商垄断价格的胁迫。2019 年 2 月，加州大学在终止与爱思唯尔的协议后，于同年 4 月，同剑桥大学出版社签署了美国史上第一个开放获取出版协议。但我国至今还没有能够有效应对涨价的完整的文献资源保障体系方案。重构文献资源保障体系长路漫漫，把握好重构之路需要对过往已有的研究进行宏观上的把握。通过对过往研究的梳理，了解文献资源保障体系这一框架之下具有哪些方面的研究主题和工作内容，有利于为重构文献资源保障体系提供指导借鉴，有利于改变近几年被计算机技术牵着鼻子被动向前的发展局面，从而以历史为指针，以新兴技术为滚轮，构建起自给自足的、能够与国内外数据库商相抗衡的文献资源保障体系。

2 LDA 模型与研究设计

2.1 LDA 模型介绍

为探究过往文献资源保障体系相关文献的研究主题，需要对已发表的相关文献主题演化趋势进行研究。而一篇文章的关键词有的代表研究问题，有的代表研究方法，有的代表研究对象，因此仅从关键词入手不利于对文献主题进行识别 [2]。目前既有研究大多采用主题模型的方法挖掘主题和探究主题演化。其中最简单的是词频 - 逆文档频率 (Term Frequency- Inverse Document Frequency，TF-IDF)，将文档集表示成以文档为行、以单词为列的矩阵，该矩阵的值与某一词在特定文档中的频率成正比，与其在多个文档中的频率成反比。TF-IDF 容易出现矩阵稀疏的情况，即只是从词频的角度而非以语义的形式表示文档 [3]，还容易低估在一个类中高频出现的却能够代表这个类的主题的词 [4]，因此需要不断调整 TF-IDF 的各项参数以适应实际需求 [5]。故本文以 LDA（Latent Dirichlet allocation）主题模型为基础，对历年文献资源保障体系相关文献进行主题识别。LDA 即隐含狄利克雷分布，是基于“文档 - 主题 - 词”的三层贝叶斯概率模型 [6]。具体的联合概率公式为：

$$P(\theta,z,w\,|\,\alpha,\beta)=P(\theta|\alpha)\prod_{n=1}^{N}P(z_n|\theta)\ P(w_n|z_n,\beta) \quad \text{公式 (1)}$$

其中，θ 表示主题分布，α 是主题分布 θ 的先验分布 (即 Dirichlet 分布) 参数，β 是关键词分布的先验分布参数，z 表示模型生成的主题，w 表示模型最终生成的关键词，N 表示文档的词语数量，M 表示文档数量，三层概率模型如图 1 所示：

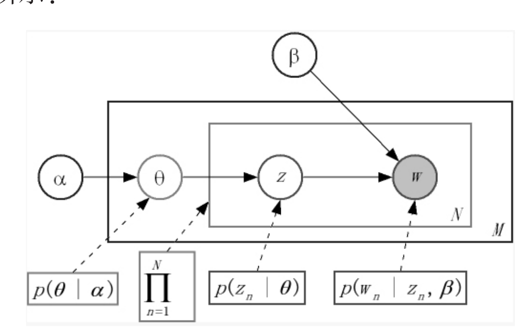


图 1 LDA 三层概率模型

2.2 研究设计

此前较少有学者对我国文献资源保障体系相关文献进行主题演化分析，本文主要利用 LDA 模型对相关文献进行主题识别，实现 LDA 模型对文献资源保障体系相关文献的应用。进一步绘制关键词共现网络和主题演化桑基图，从宏观数量层面和微观时间线层面进行演化分析，主题识别流程具体分为 4 个模块，如图 2 所示：

知识管理论坛，2021（6）:303-314
DOI: 10.13266/j.issn.2095-5472.2021.029

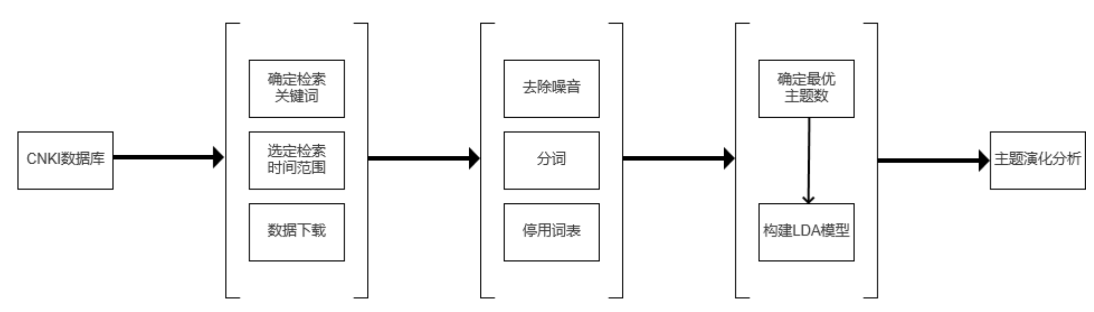


图 2 主题识别流程

（1）数据库选择和数据搜集。本文的研究对象是国内文献资源保障体系论文主题演化发展，故选择CNKI期刊全文数据库作为数据来源。检索主题词为“文献资源保障体系”“文献信息资源保障体系”“文献保障体系”和“文献资源保障”，逻辑连接词为“OR”。检索年份为 2021 年之前的所有相关文献。在人工去除部分不相关的文献之后，累积相关文献共计 1 429 篇，将相关文献的标题、关键词、摘要汇总作为数据源备用。

（2）数据预处理。根据以上所收集到的数据，对所有文献的标题、关键词和摘要信息进行合并，将其视为代表该文献的长文本，之后利用 Python 的 jieba 分词工具包进行中文分词。为了提高分词的效果，需要设置用户自定义词典，根据多次的分词试验结果，将“文献资源”“双一流”“大数据”等专有名词保存进自定义词典以提高分词结果的有效性。分词过程中还要添加停用词表，本文选用的是常用的中文停用词表——哈工大停用词表。最后将分词的结果进行保存，作为 LDA 模型构建的数据。

（3）LDA 建模。在用 LDA 模型进行主题识别前需要计算最优的主题数目。本文采用 Python 中的 scikit-learn 工具包中的 K-means 算法，通过计算分词文本的簇内误差平方和系数（Distortions）和轮廓系数（Silhouette）来确定最优聚类数 k，其中 Distortions 系数越小越好，Silhouette 系数越大越好，结果如图 3、图 4 所示：

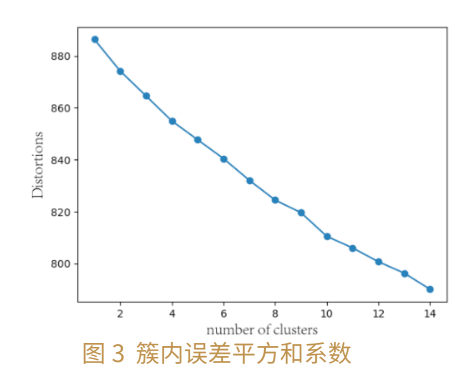


图 3 簇内误差平方和系数

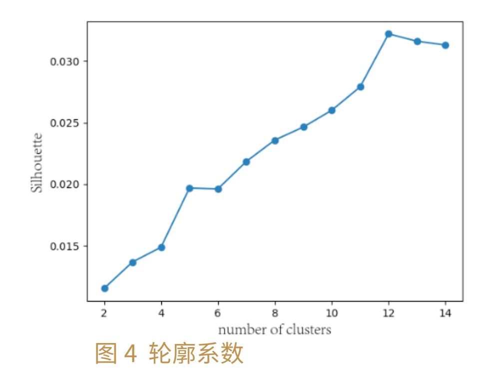


图 4 轮廓系数

由图 3、图 4 可知，综合考虑簇内误差平方和系数（Distortions）和轮廓系数（Silhouette），选择 12 个聚类数较为合适。笔者在后续的 LDA 建模中将聚类主题数设置为 12，α 和 β 均保持 python 库中的默认值。由于本文将标题、摘要和关键词统一视为一段长文本，故在此各权重一致。

要想在 LDA 模型聚类结果中探索不同主题的演化路径，除了结合文献发表时间这一自然属性之外，还要通过计算文本相似度、设定一

定的阈值来确定具有较高相关性的文本主题，进而判定为演化关系，以形成该类主题的演化路径。本文采用计算余弦相似度的方法来衡量不同年份下同一聚类内部文本的相似度，从而确定主题间的演化路径。余弦相似度的计算公式如下：

$$similarity = \cos(\theta) = \frac{A*B}{\|A\|*\|B\|} = \frac{\sum_{i=1}^{n}(A_i*B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}*\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad \text{公式 (2)}$$

余弦相似度的取值范围在（0，1）之间，取值越大说明两文本越相关，由于文摘部分内容较长，为避免数据过于稀疏，将相似度指标设置在 0.1，即两文本相似度大于 0.1 时，可认定为具有演化关系，并结合文本发表年份绘制主题演化路径。

（4）主题结果分析。结合文献资源保障体系相关文献的数量分布和 LDA 模型聚类的主题结果，进行进一步的深入分析，一方面从宏观的数量层面探究我国文献资源保障体系的相关文献数量的变化，另一方面从微观的主题层面探究我国文献资源保障体系相关文献的主题演化路径。

3 文献资源保障体系主题结果及演化分析

3.1 LDA 模型主题识别结果分析

根据历年发表的相关文献数量，绘制逐年折线图（图 5）。在 1983 年，我国颁布了《中华人民共和国国家标准·文献著录总则》(GB3792.1-83)，该文件将“文献”定义为“记录有知识的一切载体”。这一概念的提出，使得学术界对于“文献”的内涵与外延有了较为统一的定论。渐渐地，“图书”“藏书”这两个名词也都逐渐被涵盖在“文献”这一定义之下。根据检索结果，从 1984 年起，陆续有文献资源保障体系相关的文献发表，与之伴随的正是图书馆职能转变的探讨：将图书馆藏书从收藏化为利用，更好地为社会各界的文献资源需求提供保障。2000 年前后，相关文献的数量开始快速上升，而在 2010 年往后，相关文献的数量开始呈现下降的趋势。

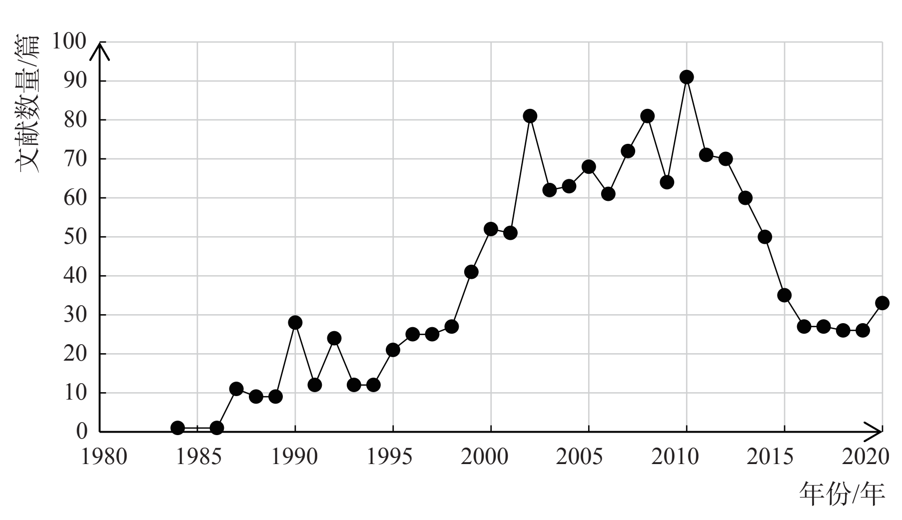


图 5 相关文献数量折线图

表 1 为 LDA 主题模型识别结果，从同一主题下的词项中，选取概率较高且具有主题意义的词项，并归纳出相应的主题标识来代表该主题。由表 1 可知，信息资源评价、文献资源建设、文献资源共享、数字图书馆等一系列与文献资源保障体系相关的主题被识别出来。根据相关文献发表数量的逐年变化，将相关的文献数据划分为三个部分，分别为 I 时期 (1984-1999 年)、

知识管理论坛，2021（6）:303-314
DOI: 10.13266/j.issn.2095-5472.2021.029

Ⅱ时期 (2000-2010 年) 和Ⅲ时期 (2011-2020 年)，并利用 CiteSpace 可视化软件绘制相关文献的关键词共现网络（图 6- 图 8），展现相关主题词之间的联系。

表 1 LDA 模型主题识别词项

主题编号	主题标识	词项					
Topic 1	信息资源评价	评价	指标体系	专家	用户	文献计量	绩效
Topic 2	特色文献资源	特种文献	地方	西部	少数民族	小语种	特色
Topic 3	图书情报机构	机构设置	业务重组	基层图书馆	图书馆联盟	馆舍建设	总分馆
Topic 4	文献资源共享	共建共享	馆际互借	文献资源布局	资源布局	资源共建	整合
Topic 5	文献资源保障系统	文献资源保障体系	CALIS	NSTL	BALIS	省级	国际
Topic 6	高校图书馆	高校图书馆	重点学科	资料室	学科导航	教学	研究生
Topic 7	数字图书馆	数字文献	数字资源	图书馆自动化	元数据	电子图书馆	智慧图书馆
Topic 8	文献资源建设	科技文献资源	标准文献	农业文献	体育文献	医学文献	农业文献
Topic 9	文献传递	文献传递	原文传递	期刊订购	查新	文传模式	费用
Topic 10	文献收藏	馆藏	经费	数据库	文献保护	修复	剔旧
Topic 11	信息服务	知识服务	参考咨询	知识经济	产权	法律保护	文献出版
Topic 12	文献组织	元数据	数据挖掘	MARC	检索	采访	文献组织

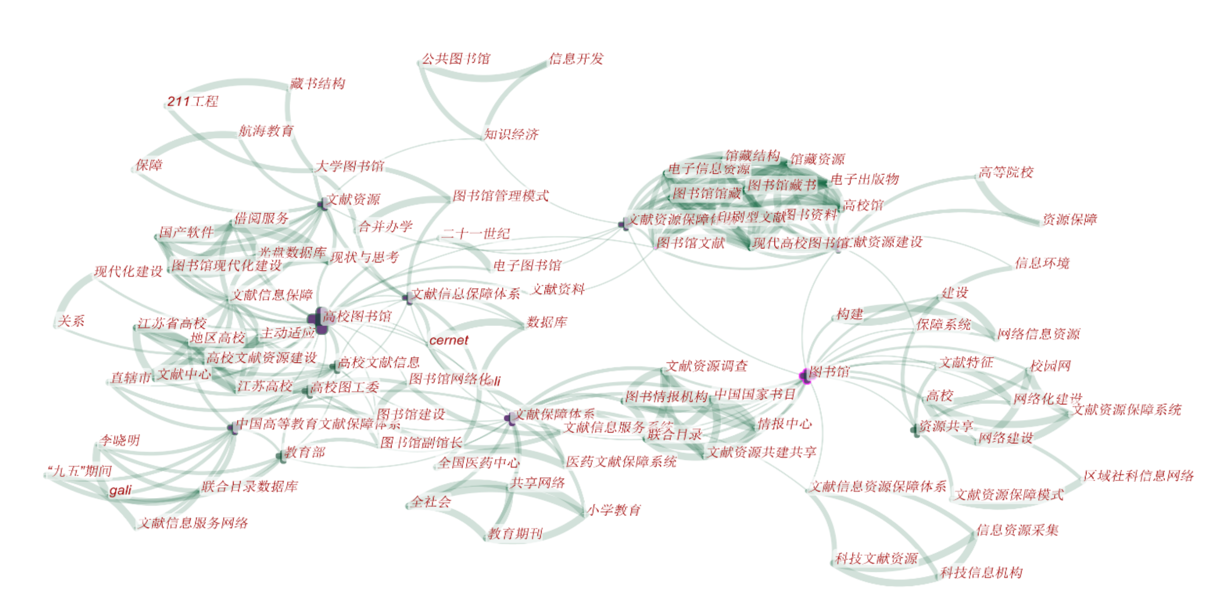


图 6 Ⅰ时期 (1984-1999 年) 关键词共现网络

Ⅰ时期 (1984-1999 年) 的主题主要集中在 Topic3 图书情报机构、Topic4 文献资源共享、Topic6 高校图书馆、Topic8 文献资源建设、Topic10 文献收藏与 Topic12 文献组织。研究内容主要是传统的图书馆等图书情报机构职能研究与新世纪的展望和规划。



图 7 II 时期 (2000-2010 年) 关键词共现网络



图 8 III时期 (2011-2020 年) 关键词共现网络

20 世纪的文献资源主要还是以纸质文献资源为主，但是数字化的文献资源也逐渐发展起来，与之相对应的是数据库的变化发展，1986 年，国家海洋局情报所首先引进国外只读光盘（CD-ROM）数据库以后，各高校图书馆和情报机构也纷纷引入只读光盘，用于课题检索，具体包括定题服务、回溯检索、专题服务、成果查新和专利审查的查新等 [7]。1992 年，由中国科技情报研究所重庆分所数据库研究中心推出的《中文科技期刊篇名数据库》（CB ISTIC/CEPC Periodicals ChinaBase）只读光盘版正式发行，系我国大陆第一张中文数据光盘。1997 年 1 月，《中国学术期刊（光盘版）》正式定期发行，是我国第一部大规模集成化学术期刊全文数据库，图书馆界将此视为我国进入数字图书馆时代的标志和里程碑。

然而，互联网的发展速度远快于光盘数据库的发展速度。在世界银行的《1998 年度世界发展报告》提出国家知识基础设施（National Knowledge Infrastructure，NKI）的概念之后，1999 年 3 月，王明亮提出要建设中国知识基础设施工程（China National Knowledge Infrastructure，CNKI）。重庆维普资讯有限公司于 2000 年建立了维普资讯网。万方数据公司在 20 世纪 90 年代初推出国内第一个资讯产品——《中国企业、公司及产品数据库》。至此，知网、维普和万方逐渐成为国内主流的三大数据服务平台。

Ⅱ 时期 (2000-2010 年) 和 Ⅲ 时期 (2011-2020 年) 的主题主要集中在 Topic8 文献资源建设、Topic4 文献资源共享、Topic6 高校图书馆、Topic5 文献资源保障系统等主题。可见 21 世纪所面临的主要挑战是建设面向新时代、面向社会各个领域的文献信息资源，逐步建立起文献资源保障体系。

文献资源保障体系是一个集文献的收集、

知识管理论坛，2021（6）:303-314
DOI: 10.13266/j.issn.2095-5472.2021.029

贮存、揭示、传递、利用等诸多功能为一体的社会系统[8]。在整个文献资源保障体系的运行模式上，肖希明认为等级结构控制的方式是构建我国文献资源保障体系模式的正确选择，在以大系统的等级结构控制为基本构架的同时，吸收其他控制方式的优点，构建一个由地区(省、市、自治区)级、区域(行政大区)级和国家级文献资源网构成的三级网络结构模式[8]。孙瑞英在此基础上提出增加建立国际级保障体系的建议[9]。

Topic5 文献资源保障系统是文献资源保障体系研究中的重要实践。作为我国最早启动的文献信息资源保障系统，“中国高等教育文献保障系统”（China Academic Library & Information System，CALIS）于 1998 年正式成立，CALIS 作为“211 工程”建设的公共服务体系之一，为各高校的重点学科发展起到了支撑保障作用，内容上涵盖了农业文献、法学文献、商业文献、医学文献、体育文献等各个不同学科领域，结构上包括了标准文献、科技文献和外文文献等不同类型的文献。除 CALIS 之外，后续开发出了多个文献资源保障系统，例如：北京地区高校图书馆文献资源保障体系（BALIS）[10]、江苏省高等教育文献保障系统（JALIS）[11]、中国高校人文社会科学文献中心（CASHL）[12]、国家科技图书文献中心（NSTL）[13] 和国家科学数字图书馆（CSDL）[14] 等。

进入Ⅲ时期 (2011-2020 年)，自 2015 年国务院发布了《统筹推进世界一流大学和一流学科建设总体方案》（简称“双一流”建设）[15] 之后，“211 工程”建设逐渐转为“双一流”建设，CALIS 以及其他文献资源保障系统的职能也发生了相应变化，主要为高校“一流学科”的文献信息需求提供保障。

3.2 主题演化分析

除了较为粗粒度地将文献主题划分为三大时期进行关键词共现分析之外，本文还根据余弦相似度，结合文献的发表时间来绘制桑基图（Sankey Diagram），进一步探究文献资源保障体系主题的演化路径。桑基图，又称为桑基能量分流图，起源于 1898 年的“蒸汽机的能源效率图”。在桑基图中，对象用元素块来表示，对象间产生能量的流动方向及联系则通过连线来表示。本文的元素块表示某一研究主题，主题之间的连线表示主题之间的演化关系，主题元素块后面的括号中标注了首次出现该主题的年份。为使主题的演化路径更清晰，重复出现的主题词在后续的演化路径中将不再表现出来。

由图 9 可知，文献资源保障体系的主要研究内容形成时间都比较早，到 2000 年左右，文献资源保障体系的相关文献主题已基本涵盖，后续具体探讨的是新世纪新环境之下，不同建设领域其内容和方式方法上的更新与完善。下文主要对各个主题内部的演化进行分析。

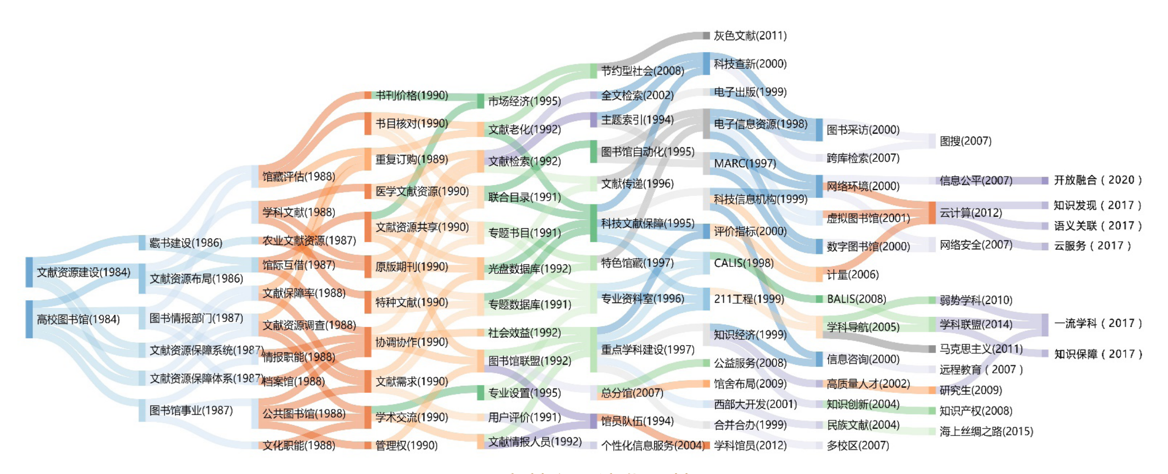


图 9 文献主题演化桑基图

（1）Topic4 文献资源共享是文献资源保障体系的主要探究主题。1973 年，国际图书馆协会联合会（国际图联，International Federation of Library Associations and Institutions，IFLA）首先提出了 UAP（Universal Availability of Publications）的概念，即“国际图书馆资源共享”。该理念引入国内后，引起了国内学者们的高度赞同。远征就指出，实现文献资源共享，有利于高校图书馆摆脱“自给自足”的自然经济状况，可以缓解图书经费不足、管理人员欠缺、编目能力有限、藏书空间饱和等一系列问题 [16]。其中，馆际互借是实现文献资源共享的重要途径与手段，我国最早提出馆际互借大约是在 1939 年，当时民国政府教育部颁布了《修正图书馆规程》和《图书馆工作大纲》，文中有提到“馆际互借与邮寄” [17]，某种程度上说，这是文献资源共享的开端。但受限于当时国内时代背景，文献资源共享一直未能得到良好的发展。到了 20 世纪末，由于信息技术的发展，众多学者开始倡导图书馆自动化 [18-19] 和数字图书馆 [20-21]，实体书的馆际互借也慢慢发展为线上的更具有广泛意义上的资源共享，从而演化出了一系列与互联网技术发展紧密相关的研究主题，例如编目标准化、网络建设、信息资源网络化等问题。在这期间，也有学者倡议建设文献资源共享服务网络中心，或是采用有偿共享的模式 [22]，但后续的相关研究成果较为有限。在 2006 年 8 月，Google 首席执行官埃里克·施密特（E. Schmidt）在搜索引擎大会（SESSanJose2006）上首次提出“云计算”（Cloud Computing）这一概念 [23]。云计算本质上是一种全新的网络应用概念，使用者可以随时获取“云”上的资源，按需求量使用 [24]，这一概念的提出，给文献资源的共享模式带来了新的思路。

与此同时，中国作为农业大国，农业文献信息资源的共享对于国家的发展进步同样具有战略意义。全国农业文献资源共建共享的最终目的是为全国的农业教学、科研、生产和经营所需要的资源支持和服务提供保障，进而促进我国农业的突破性发展 [25]。为提高西部地区的经济和社会发展水平，巩固国防，国务院于 2000 年 1 月成立了西部地区开发领导小组。此前，已有学者针对西部地区的农业文献资源共建共享提出相应的举措 [26]。而相关文献保障体系的建立，也将有利于图书馆为西部大开发建言献策，同时为西部地区做好文献保障 [27]。随着技术的发展，目前，中国农业科学院信息化服务网也已经上线，通过一系列信息化服务，打造智慧农科协同平台。2013 年 9 月和 10 月由中国国家主席习近平分别提出的建设“新丝绸之路经济带”和“21 世纪海上丝绸之路”的合作倡议，打开了我国西部地区和沿海地区的大门。做好相关历史文献的梳理和保障工作，不仅有利于申报世界文化遗产，推动特色文献资源的建设，还可以加强对周边国家地域文化的研究，甚至在一定程度上缓解边境或沿海岛屿的争端 [28-29]。

（2）Topic9 文献传递、Topic10 文献收藏、Topic11 信息服务和 Topic12 文献组织均是 Topic8 文献资源建设过程中演化出来更加细分的主题。文献传递与收藏是图书馆的基本职能。高校图书馆馆藏文献资源是高校重点学科建设的重要保障。由于不同省市的经济发展情况和当地特色文献的数量不同，文献资源布局的情况均有所不同，对相关文献进行采购与收藏之前，需要对文献资源布局进行充分的调研，然后对缺少的有需求的文献资源进行采购并收藏。文献收藏与传递除了关注数据库技术的演化发展之外，还涉及多重备份与适时迁移、开放描述方式、模拟环境与环境封装、数据恢复与数据考古、技术框架与整体解决方案、标准化技术等多个方面 [30]。

为了更好地收藏与传递文献资源，需要对其进行有效的描述。元数据是文献资源组织中信息描述的重要部分，元数据不但在数字资源著录方面具有重要的作用，也是使得图书馆走向自动化的关键技术。MARC（Machine-Readable Catalogue，机器可读目录）与 Dublin

DOI: 10.13266/j.issn.2095-5472.2021.029

Core（都柏林核心集）两种元数据发展较为成熟，并且在图书情报界得到广泛的认可。1965年，由美国国会图书馆研发的MARC(后来称之为MARC Ⅰ)，代表了机读目录的初步成果，后在英美合作之下，MARC Ⅱ于1968年问世。我国有关部门于1991年在UNIMARC的基础上加上特定字段，编制了《中国机读目录通讯格式》(CNMARC)，并多次修订。1995年，OCLC和NCSA联合召开了第一次都柏林核心集会议，最终确立了包含15个核心元素的核心集。由于MARC在粒度、语言和可扩展性方面具有一定的局限，美国国会图书馆(Library of Congress，LC)于2011年5月提出了书目框架模型（Bibframe），力求大大整合现有的书目资源，但其如何适应中文的编目环境还有待深入研究。新时代互联网环境的迅速发展将持续推动文献信息资源组织方法及理念的创新和改革，文献信息资源组织将朝着跨学科融合、智能语义组织以及信息方法一体化等方向快速发展[31]。

文献资源建设的最终目的依然是服务用户、服务读者。1995年5月，江泽民同志在全国科技大会上的讲话中提出了要实施科教兴国的战略。这促使图书馆从信息服务走向知识服务，通过知识服务助推科教兴国战略的实施。知识服务是指从各种显性和隐性信息资源中，针对人们的需要将知识提炼出来、传输出去的过程[32]。知识服务正是以文献信息资源建设为基础的高级阶段的信息服务。要想充分开展知识服务，需要深入挖掘用户的知识需求，通过智慧的手段使显性知识增值，使隐性知识可以被传递和接收，从而提供个性化信息服务[33]，这些都需要知识挖掘、知识组织、知识开发和知识服务人员素养等多方面的提升[34]。与此同时，在线信息服务提供商和大型出版商逐渐开始了语义网应用实验，产生了语义出版这种新的出版形态，语义出版将文献资源从一个孤立、静止的知识包变成了嵌入在相互关联和相互作用的知识体系中的知识工具[35]。语义出版一方面帮助用户发现或验证新知识，另一方面能使出版机构获得新的利润回报和盈利空间。长远看来，知识服务和语义出版仍将是信息服务领域的一片红海。

（3）Topic1 信息资源评价。在宏观层面上，建设文献资源保障体系离不开高层次的宏观调控机构，相应的政策、法规和标准以及社会各界的力量[36]。微观层面上，为了保证更好地建设文献资源保障体系，需要对相关的主体和客体进行评价评估。索传军等将评价主体分为个体和机构两类，评价者个体是指来自于不同领域的专家学者，而评价机构则包括经营性机构、服务性机构、学术性机构等[37]，再根据不同的评价客体形成不同的评价体系，例如期刊评价体系、馆藏文献资源评价体系、数字文献资源评价体系等。安月英构建了一个二级的馆藏资源评价体系，其中一级指标包括资源内容、检索系统、经济性和存储系统，二级指标包括馆藏资源保障能力、权威性、时效性、规范性、检索功能、检索效果、易用性、成本、使用情况、存储系统的效率和安全性[38]。马海群等从信息源内容、信息源组织、信息源性能、其他指标这四大层面构建了一套含有16个指标的网络信息资源评价体系[39]。而期刊评价体系的指标包括但不限于总下载量、影响因子、5年影响因子、他引影响因子、平均引文数、Web即年下载率、即年指标、综合总被引、可被引文献量、引用期刊数、被引期刊数、等各类指标[40]。

与其同时，高校的学科评估一部分也是对相关学科的文献资源进行评估。通常，我国高校院系可根据科研工作和教学需要，自主购买中、外文文献，这种自主采购的方式针对性和专业性很强[41]。但也有高校图书馆文献资源采购，在以满足师生的阅读需求的前提之下，兼顾院校的重点学科的发展，打造特色馆藏和重点学科馆藏。随着“双一流”等一系列工程的实施，各个高校在办学的过程中还出现了高校合并、多校区办学的情况，这使得高校图书馆在政治思想工作、机构设置和人事管理、规章制度标准化、网络和软件更新、经费管理、资

源共享、馆藏布局等多个方面面临变革[42-43]。这一系列高校和学科的变革终将需要相关文献资源服务的配套优化。

总的来看，文献资源保障体系的各个主题是相互交织在一起一同发展的，其主题演化大致与图书馆自动化的四个发展阶段和信息技术发展的趋势保持一致。第一阶段为图书馆自动化管理集成系统发展阶段，第二阶段为图书馆在网上进行全球性、整体化的电子文献信息服务的阶段[44]，第三阶段为数字化图书馆阶段，第四阶段便是智慧图书馆阶段。在这期间，信息技术不断地更新迭代，如光盘 CD-ROM 的兴衰，从局域网到互联网，Bibframe 模型逐渐替代 MARC，云计算、大数据、物联网等一系列新技术蓬勃发展。国家层面也发布不同的政策文件，从“211 工程”到“双一流学科”，从“西部大开发”到“一带一路”，从“九五”的“金图”工程到“十四五”的网络空间命运共同体，都在不断地推进我国文献资源保障体系的发展演化。综合来看，我国文献资源保障体系的演化是在图书情报、计算机等众多学界的共同努力之下，依托先进的信息技术，不断为中国文献保障事业添砖加瓦的过程。

4 结论与讨论

本文主要基于 LDA 主题模型进行主题识别，实现了 LAD 主题模型在文献资源保障体系领域的应用。在文献资源保障体系的主题演化路径中，形成了丰富多样的主题，2000 年左右基本已包含主要的文献资源保障体系研究主题，主要可以分为文献资源层面和机构层面两大内容。从资源层面来看，研究内容包括文献资源类型与收集、文献资源组织与建设、文献资源服务与共享。从机构层面来看，高校图书馆一直是研究的主要对象，随着文献资源保障体系发展建设，全国性机构与地方基层机构建设发展迅速。这两大内容还受到“西部大开发”、“双一流”学科建设、“一带一路”等一系列宏观政策和“大数据”“云计算”“数据挖掘”等一系列新兴技术的交叉影响。

为重构文献资源保障体系，同样需要从资源角度和机构角度进行相应的调整。从资源的角度来看，在文献资源类型与收集环节，部分高校的文献资源向重点学科、强势学科倾斜，忽视弱势学科、少数民族地区、非英语外文文献等资料的采购，但从体系优化的层面上看，同样要兼顾“弱势学科”和多元发展，从而带动各高校各学科的水平提升。在文献资源组织与建设环节，数字化建设是文献资源建设的大方向，要将元数据建设作为其核心，实现多渠道元数据融合、多类型元数据映射、多层次元数据识别，建立具有知识关联功能的智慧的文献组织平台。在文献资源服务与共享环节，建立以开放数据、开放获取、开放出版等为手段的数字资源开放生态新模式的同时，应充分重视用户需求与用户价值，通过抓取分析用户的主观特征、行为数据、偏好数据、意见反馈，构建用户画像，为用户提供细粒度的个性化的资源服务。

从机构的角度来看，虽然多年来发展出各类专门专项的机构成员，但存在不同程度的冗杂、分块严重、职责重复、缺乏统一管理等问题。建立统筹协调、部门联动的文献资源保障体系的管理机制，首先需要在中央设立跨系统、跨学科、跨部门的全国统一常设机构，总揽相关文献资源的指挥与协调；其次，向下设立全国性的专家委员会与具体办事机构负责技术指导与执行；最后，依托各级学会与地方图书馆下设各系统间的地区联盟和基层组织，实现分类、分级、分工保障[45]，从而建成全国性的横跨各个领域的文献资源保障体系系统。

本文具有一定的局限：LDA 模型聚类中，个别聚类内的文献数较少，不能充分反映文献的演化规律。在进行演化路径分析的时候，后续再次出现的主题词不再作为分析展示的对象，一定程度上会忽略主题演化过程中更为微观的演进变化，后续还将继续在文献资源保障体系主题演化更为细致的方面深入研究，并探讨演

ChinaXiv合作期刊
知识管理论坛，2021（6）:303-314
DOI: 10.13266/j.issn.2095-5472.2021.029

进的机理机制和未来新主题的识别预测。

参考文献：

[1] 王翠萍 , 杨沛超 . 国家文献信息资源保障体系建设论纲 [J]. 图书馆学研究 , 2000(2): 15-17,14.
[2] 刘自强 , 王效岳 , 白如江 . 多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例 [J]. 中国图书馆学报 , 2016, 42(6): 67-84.
[3] 范云满 , 马建霞 . 利用 LDA 的领域新兴主题探测技术综述 [J]. 现代图书情报技术 , 2012(12): 58-65.
[4] 张玉芳 , 彭时名 , 吕佳 . 基于文本分类 TFIDF 方法的改进与应用 [J]. 计算机工程 , 2006(19): 76-78.
[5] 施聪莺 , 徐朝军 , 杨晓江 . TFIDF 算法研究综述 [J]. 计算机应用 , 2009, 29(S1): 167-170,180.
[6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation. Journal of machine learning research, 2003, 3: 993-1022.
[7] 夏旭 , 曾海标 . CD-ROM 数据库的引进、开发、利用现状及对策 [J]. 图书馆杂志 , 1996(3): 32-35.
[8] 肖希明 . 我国文献资源保障体系论纲 [J]. 图书馆 , 1996(6): 8-12.
[9] 孙瑞英 . 建立国家文献信息保障体系的构想 [J]. 情报科学 , 2002(7): 680-683.
[10] 黎晓 . 北京高校文献资源保障体系发展现状分析 [J]. 农业图书情报学刊 , 2014, 26(3): 16-20.
[11] 杨永厚 . 江苏省高校文献保障系统建设的回顾与思考 [J]. 大学图书馆学报 , 2002(1): 72-74,92.
[12] 李朵 . 中国高校人文社会科学文献中心 (CASHL) 网络服务系统现状与发展 [J]. 大学图书馆学报 , 2005(3): 27-29.
[13] 袁海波 , 孟连生 . 网络环境下信息资源共建共享的实践——兼述国家科技图书文献中心的建设与发展 [J]. 情报学报 , 2002(1): 57-62.
[14] 张晓林 . 国家科学数字图书馆及其建设进展 [J]. 中国科学院院刊 , 2005(4): 344-346,343.
[15] 朱丽莉 . “双一流”建设背景下高校图书馆文献资源建设策略探讨 [J]. 图书情报导刊 , 2018, 3(8): 7-11.
[16] 远征 . 实现资源共享对高校图书馆的现实意义 [J]. 大学图书馆通讯 , 1987(4): 10-13,19.
[17] 崔慕岳 , 代根兴 . 论馆际互借 [J]. 河南图书馆学刊 , 1990(3): 54-57.
[18] 夏旭 . 资源共享发展的一大趋势——光盘技术与通信技术的整合 [J]. 大学图书馆学报 , 1995(5): 40-42.
[19] 黄建年 . MARC 数据与图书馆 [J]. 津图学刊 , 1997(4): 28-34.
[20] 曹作华 . 论网络化、数字化与高校图书馆馆藏建设策略的转化 [J]. 情报科学 , 2002(1): 16-18.
[21] 王元如 , 宁圣红 . 数字图书馆和文献信息资源共建共享 [J]. 现代情报 , 2000(6): 15-16.
[22] 郭晔 . 浅谈创建文献资源有偿共享体系 [J]. 宁德师专学报 (哲学社会科学版), 2004(2): 82-83,86.
[23] 许子明 , 田杨锋 . 云计算的发展历史及其应用 [J]. 信息记录材料 , 2018, 19(8): 66-67.
[24] 罗晓慧 . 浅谈云计算的发展 [J]. 电子世界 , 2019(8): 104.
[25] 宛章齐 . 试论全国农业文献资源的共建与共享 [J]. 农业图书情报学刊 , 1997(1): 21-22.
[26] 王子玉 . 关于西北五省农业图书馆文献资源共享建设的构想 [J]. 甘肃科技 , 1998(6): 2-3.
[27] 黄权才 . 图书馆参与西部大开发的策略 [J]. 图书馆界 , 2001(4): 1-6.
[28] 陈彬强 . 海上丝绸之路文献资源保障体系建设 [J]. 图书馆建设 , 2015(5): 88-92.
[29] 周纯 , 冯彩芬 , 马翠嫦 . 中国周边区域研究文献的需求与保障——以中山大学为例 [J]. 大学图书馆学报 , 2016, 34(5): 73-77,83.
[30] 王伟 . 数字资源长期保存的技术研究 [J]. 情报科学 , 2012, 30(11): 1751-1754.
[31] 魏敏 . 信息组织 4.0: 变革历程和未来图景 [J]. 国家图书馆学刊 , 2018, 27(1): 78-85.
[32] 田红梅 . 试论图书馆从信息服务走向知识服务 [J]. 情报理论与实践 , 2003(4): 312-314.
[33] 易明 , 王学东 , 邓卫华 . 基于社会网络分析的社会化标签网络分析与个性化信息服务研究 [J]. 中国图书馆学报 , 2010, 36(2): 107-114.
[34] 赵萍 , 马江宝 . 论图书馆的知识服务及其实现 [J]. 图书馆学研究 , 2005(8): 32-35.
[35] 魏蕊 , 初景利 . 学术图书开放获取与美国大学图书馆出版服务 [J]. 大学图书馆学报 , 2014, 32(3): 17-22.
[36] 汪涛 , 肖希明 . 新信息环境下的文献资源保障系统建设 [J]. 图书与情报 , 1999(1): 33-36.
[37] 索传军 , 吴启琳 . 国内外网络信息资源评价研究进展 [J]. 现代图书情报技术 , 2006(8): 55-59,93.
[38] 安月英 . 基于层次分析法的数字馆藏评价指标体系 [J]. 图书馆 , 2008(4): 82-84.
[39] 马海群 , 吕红 . 网络信息资源评价指标体系及其动态模糊评价模型构建研究 [J]. 情报科学 , 2011, 29(2): 166-171.
[40] 陈小山 , 陈国福 , 张瑞 . 基于因子分析和 SEM 模型的期刊评价指标结构关系研究 [J]. 情报科学 , 2016, 34(10): 61-64,71.
[41] 唐定海 . 院系自采文献管理初探 [J]. 图书馆建设 ,

2009(1): 52-54.
[42] 李家清 . 合并高校图书馆面临的问题及对策 [J]. 大学图书馆学报 , 2001(S1): 10-12,30.
[43] 杨树雨 . 论如何建立多校园的效益型图书馆 [J]. 情报资料工作 , 2001(2): 54-56.
[44] 杨宗英 , 郑巧英 , 夏佩农 . 图书馆自动化发展的新阶段 [J]. 大学图书馆学报 , 1997(3): 1-5.
[45] 朱泽 , 李玉海 , 王常珏 , 等 . 重构之路，我国数字资源保障体系的发展与未来——“2021 年全国数字资源保障体系重构学术研讨会”评述 [J]. 数字图书馆论坛 , 2021(6): 30-35.

作者贡献说明：
田栩冉：进行数据分析、论文撰写及修订；
马笑笑：进行数据分析、论文撰写及修订；
李玉海：负责论文选题，提出论文框架，进行论文修订与定稿。

Identification and Evolution Analysis of Literature Themes in Literature Resource Guarantee System in China

Tian Xuran[1] Ma Xiaoxiao[1] Li Yuhai[1,2]
[1]School of Information Management, Central China Normal University, Wuhan 430079
[2]China Library Innovation and Development Research Center, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] By analyzing the path of thematic evolution of literature resource guarantee system in China, this paper provides references for reconstructing literature resource guarantee system in China. [**Method/process**] This paper used LDA model to cluster the collected documents. Firstly, it used the time period as the basis of division, drew the co-occurrence network of topic words, and explored the interactions between topics. Then the paper determined the evolution path within each subject by similarity calculation and showed it in the form of visualization by drawing Sankey diagram. [**Result/conclusion**] The study finds that all the relevant themes of literature resource guarantee system in China have basically appeared around 2000. The topic mainly includes two categories of resources and institutions, which are greatly influenced by computer technology and national policy. The corresponding countermeasures and suggestions are given for these two categories of topics.

Keywords: literature resource guarantee system LDA thematic identification thematic evolution